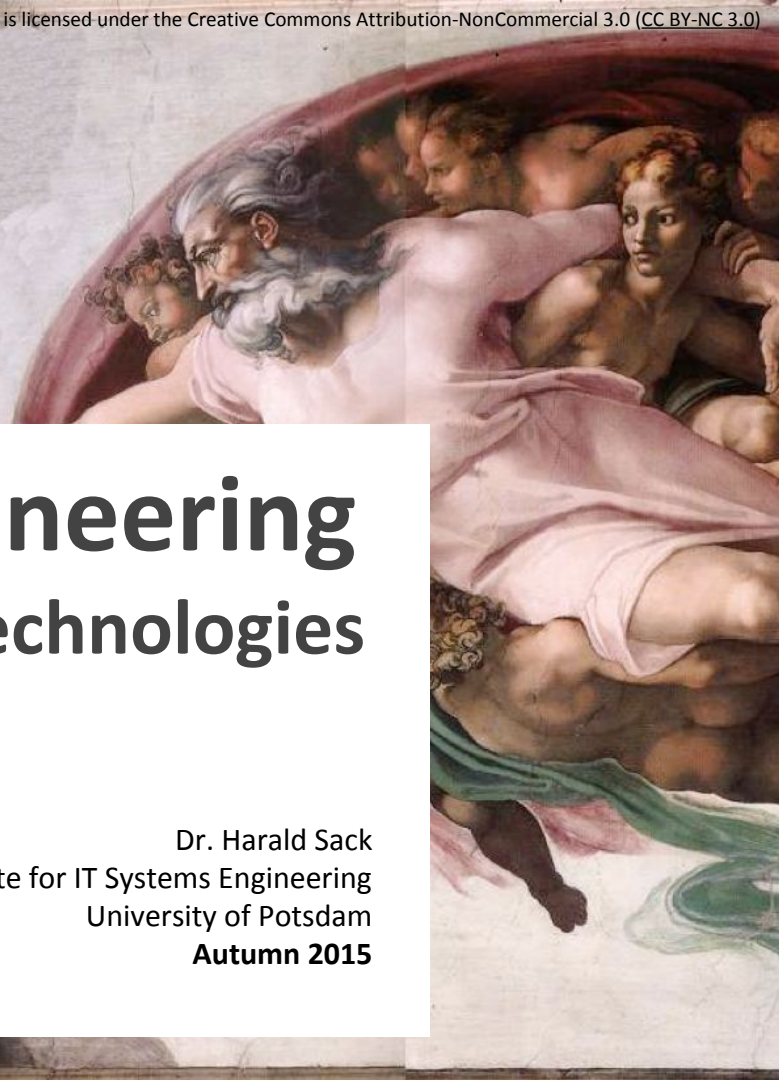
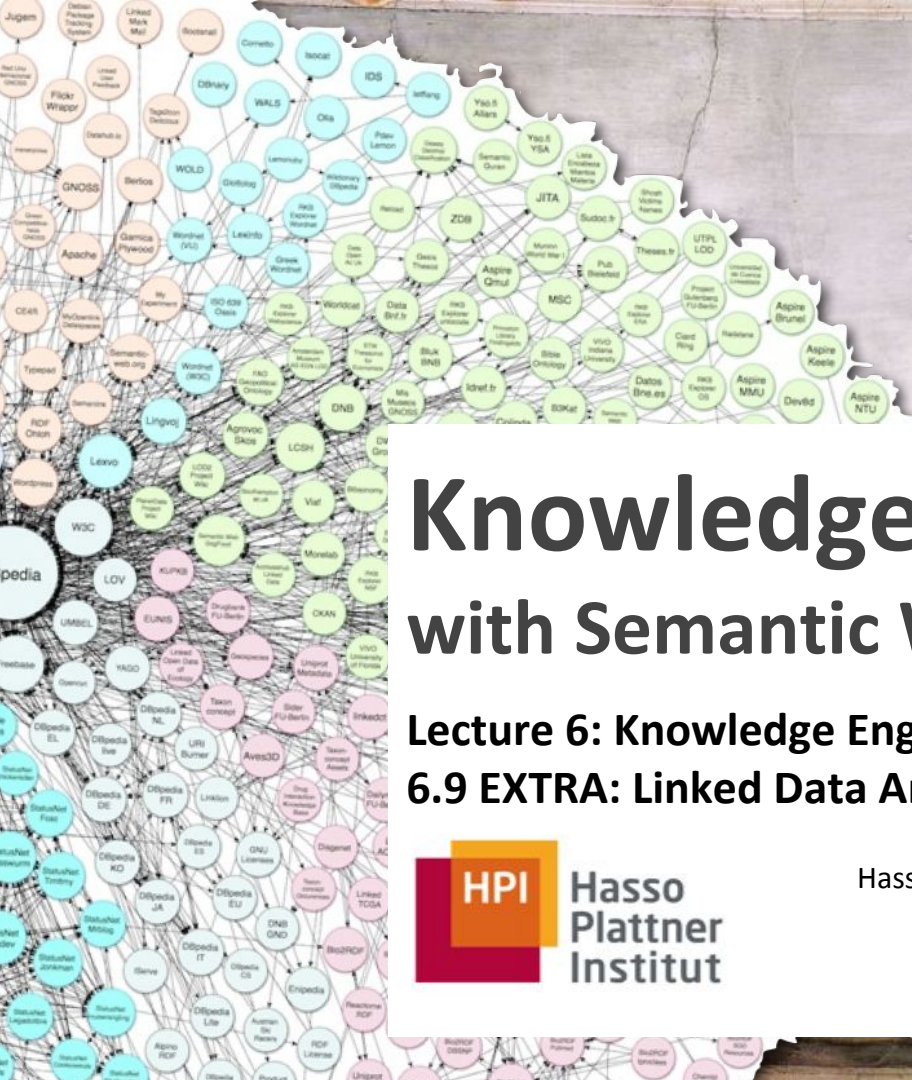




Knowledge Engineering with Semantic Web Technologies

Lecture 6: Knowledge Engineering 6.9 EXTRA: Linked Data Analytics



Dr. Harald Sack
Hasso Plattner Institute for IT Systems Engineering
University of Potsdam
Autumn 2015



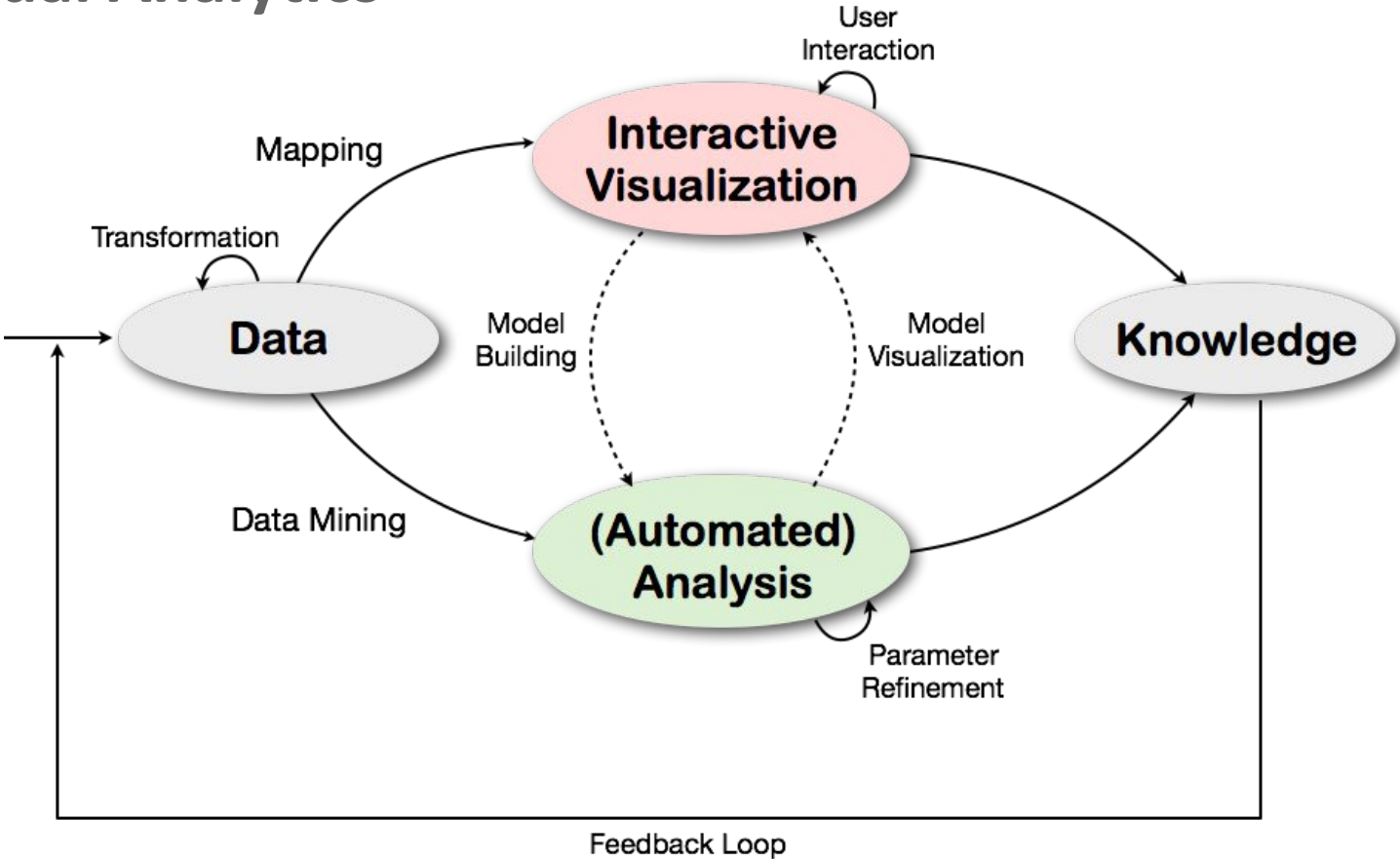
Long. List. Review of Books

A Picture tells more than a 1000 words...

- Pictures have been used to convey information long before the development of writing
- A single picture can be processed (“understood”) much faster than a text page
- Human perception is processing in **parallel**, text analysis is limited by the **sequential** process of reading

- **Information Visualization** is the study of (interactive) visual representations of abstract data to reinforce human cognition.
- **Information graphics** or **infographics** are graphic visual representations of information, data or knowledge intended to present information quickly and clearly
 - a static form of information visualization
 - aims to emphasize specific findings gained from the visualized data
 - **Mandatory precondition:** Data Analysis

- **Data Analysis** is a fundamental iterative process:
 1. Formulation and execution of a query
 2. Analysis of the results
 3. Formulation of a consecutive query based on the achieved results
- **Goals of Data Analysis:**
 - maximize understanding of analyzed data
 - uncover hidden structures/patterns
 - extraction of important variables
 - detection of anomalies and outliers
 - testing of hypotheses
 - development of a simple model



Interactive Data Analysis with Linked Data

- Available Toolset:
 - Data: DBpedia SPARQL endpoint
- Query language:
 - SPARQL
- simple statistics and visualization:
 - R

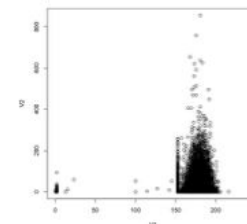


```
Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)
http://dbpedia.org

Query Text
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?type COUNT(DISTINCT ?person)
WHERE {
  ?person rdf:type dbpedia-owl:SoccerPlayer .
  ?person rdf:type ?type FILTER (!regex(?type,"player|football|people|soccer|expatriate|emigrant|citizen","i"))
}
GROUP BY ?type
HAVING COUNT(DISTINCT ?person)<20000
ORDER by DESC(COUNT(DISTINCT ?person))
```



Example from DBpedia:

- How Do I determine the importance of an entity within the Linked Data Graph

(1) InDegree

- simple Link Popularity
- can be computed via SPARQL

(2) PageRank

- more sophisticated graph measure for the importance of a linked resource
- available at public SPARQL endpoint from separate graph:

http://people.aifb.kit.edu/ath/#DBpedia_PageRank

- **Page Rank at DBpedia**
 - First example: Determine the most important astronauts

```
PREFIX vrank:<http://purl.org/voc/vrank#>

SELECT ?astronaut ?rank
FROM <http://dbpedia.org>
FROM <http://people.aifb.kit.edu/ath/#DBpedia_PageRank>
WHERE {
    ?astronaut rdf:type dbo:Astronaut .
    ?astronaut vrank:hasRank/vrank:rankValue ?rank .
} ORDER by DESC(?rank)
```

Linked Data Analytics

- now save your result locally as **csv** for further analysis
- -> astronaut.csv

[Link to the SPARQL query](#)

astronaut	rank
http://dbpedia.org/resource/Yuri_Gagarin	26.3922
http://dbpedia.org/resource/Neil_Armstrong	22.4964
http://dbpedia.org/resource/Buzz_Aldrin	15.8454
http://dbpedia.org/resource/Alan_Shepard	11.0699
http://dbpedia.org/resource/Alexey_Leonov	10.735
http://dbpedia.org/resource/Gus_Grissom	7.77648
http://dbpedia.org/resource/Valentina_Tereshkova	7.76035
http://dbpedia.org/resource/Jim_Lovell	7.11609
http://dbpedia.org/resource/Wally_Schirra	5.74465
http://dbpedia.org/resource/Eugene_Cernan	5.6846
http://dbpedia.org/resource/Yang_Liwei	5.38282
http://dbpedia.org/resource/Gordon_Cooper	5.37586
http://dbpedia.org/resource/Charles_Bolden	5.00321
http://dbpedia.org/resource/Michael_Collins_(astronaut)	4.9299
http://dbpedia.org/resource/John_Young_(astronaut)	4.90272
http://dbpedia.org/resource/Edward_Higgins_White	4.89431
http://dbpedia.org/resource/David_Scott	4.62461
http://dbpedia.org/resource/Deke_Slayton	4.4333
http://dbpedia.org/resource/Frank_Borman	4.3165
http://dbpedia.org/resource/Pete_Conrad	4.15713
http://dbpedia.org/resource/Christa_McAuliffe	4.12509
http://dbpedia.org/resource/Scott_Carpenter	4.00116
http://dbpedia.org/resource/William_Anders	3.64431
http://dbpedia.org/resource/Alan_Bean	3.45787
http://dbpedia.org/resource/Sergei_Krikalev	3.33627
http://dbpedia.org/resource/Leroy_Chiao	3.30047

Linked Data Analytics

- First example: Determine the most important astronauts
- Analysis via R
 - read data:

```
astronauts <- read.csv("astronaut.csv", header=FALSE)
```

- summarize data:

```
summary(astronauts)
```

```
[> astronauts <- read.csv("astronaut.csv", header=FALSE)
[>
[> summary(astronauts)
      V1
Min.   : 0.1500
1st Qu.: 0.6195
Median : 0.8873
Mean    : 1.2918
3rd Qu.: 1.3836
Max.    :26.3922
> █
```

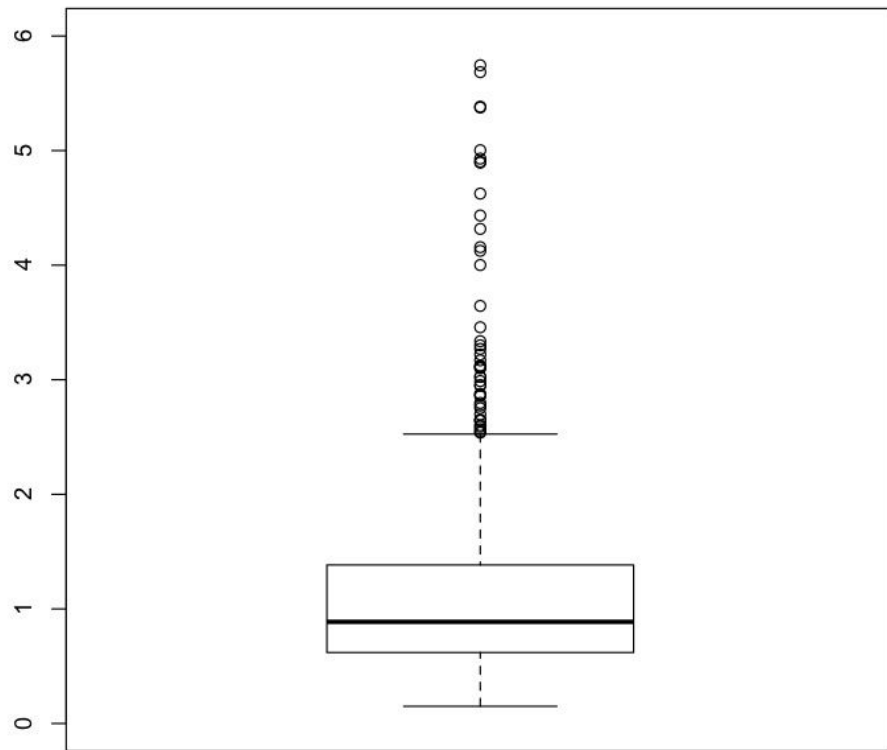


Linked Data Analytics

- First example: Determine the most important astronauts
- Analysis via R
 - analyze data via boxplot:

```
boxplot(astronauts)
```

```
[> astronauts <- read.csv("astronaut.csv", header=FALSE)
[>
[> summary(astronauts)
      V1
Min.   : 0.1500
1st Qu.: 0.6195
Median : 0.8873
Mean   : 1.2918
3rd Qu.: 1.3836
Max.   :26.3922
[> boxplot(astronauts)
[> boxplot(astronauts, ylim=c(0,10))
[> boxplot(astronauts, ylim=c(0,6))
> █
```

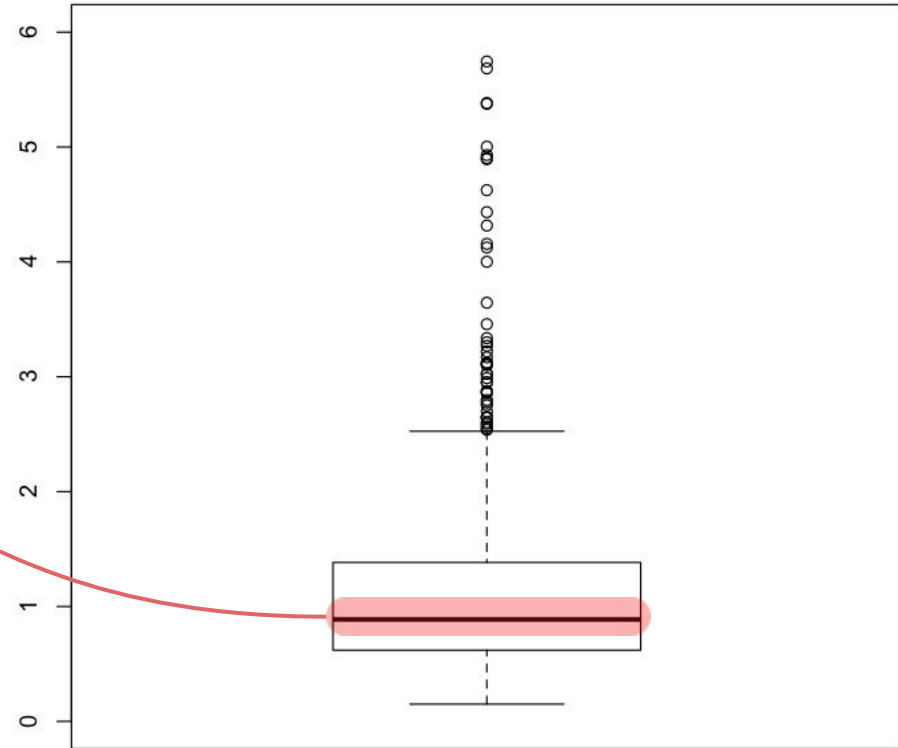


Linked Data Analytics

- First example: Determine the most important astronauts
- Analysis via R
 - analyze data via boxplot:

```
boxplot(astronauts)
```

```
[> astronauts <- read.csv("astronaut.csv", header=FALSE)
>
> summary(astronauts)
  V1
Min. : 0.1500
1st Qu.: 0.6195
Median : 0.8873
Mean : 1.2918
3rd Qu.: 1.3836
Max. : 26.3922
> boxplot(astronauts)
> █
```

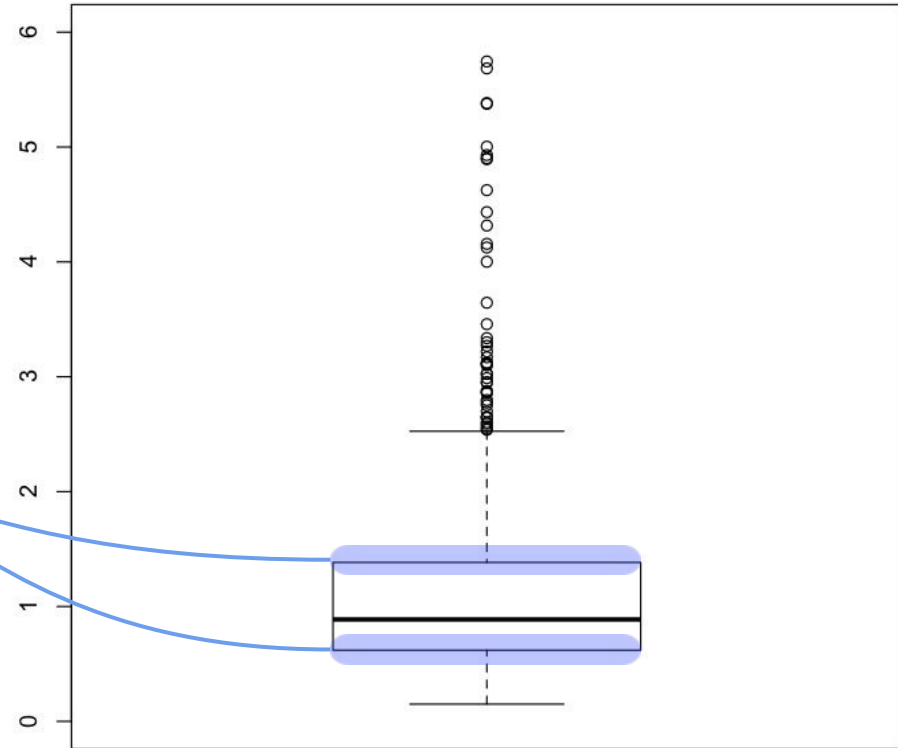


Linked Data Analytics

- First example: Determine the most important astronauts
- Analysis via R
 - analyze data via boxplot:

```
boxplot(astronauts)
```

```
> astronauts <- read.csv("astronaut.csv", header=FALSE)
>
> summary(astronauts)
      V1
Min.   : 0.1500
1st Qu.: 0.6195
Median : 0.8873
Mean   : 1.2918
3rd Qu.: 1.3836
Max.   :26.3922
> boxplot(astronauts)
> █
```



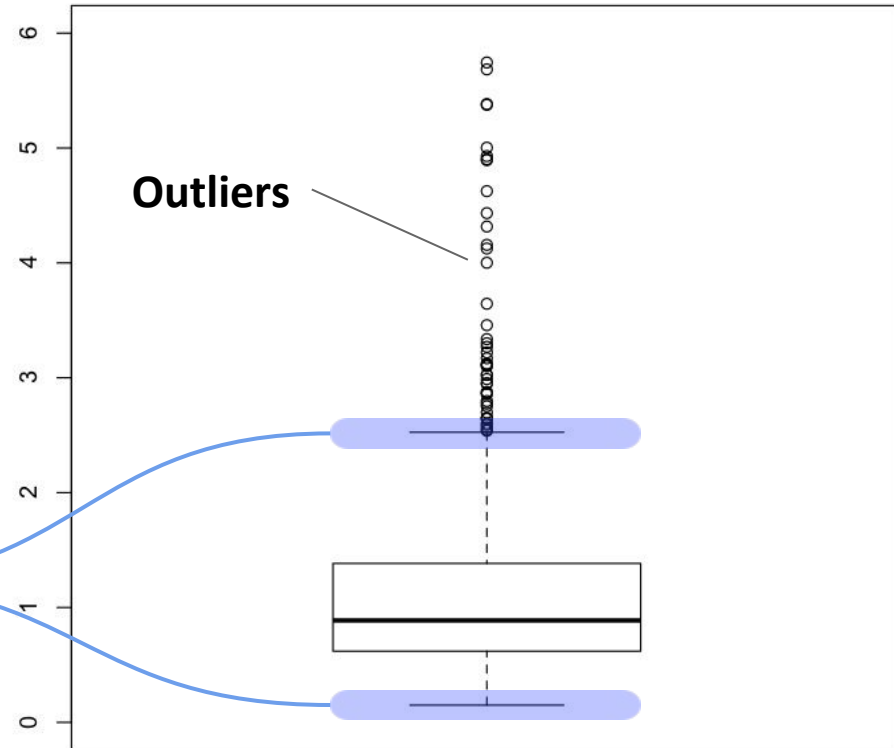
Linked Data Analytics

- First example: Determine the most important astronauts
- Analysis via R
 - analyze data via boxplot:

```
boxplot(astronauts)
```

```
[> astronauts <- read.csv("astronaut.csv", header=FALSE)
>
> summary(astronauts)
      V1
Min.   : 0.1500
1st Qu.: 0.6195
Median : 0.8873
Mean   : 1.2918
3rd Qu.: 1.3836
Max.   :26.3922
> boxplot(astronauts)
> █
```

Whiskers: $IQR \times 1.5 = (Q_3 - Q_1) \times 1.5$



- 2nd example: include **birthdate** and **Time in Space**

```
PREFIX vrank:<http://purl.org/voc/vrank#>

SELECT  ?astronaut ?rank ?time YEAR(xsd:date(?birthdate)) as ?year
FROM    <http://dbpedia.org>
FROM    <http://people.aifb.kit.edu/ath/#DBpedia_PageRank>
WHERE {
  ?astronaut rdf:type dbo:Astronaut ;
             dbo:timeInSpace ?time ;
             dbo:birthDate ?birthdate.
  ?astronaut vrank:hasRank/vrank:rankValue ?rank .
}
```

Linked Data Analytics

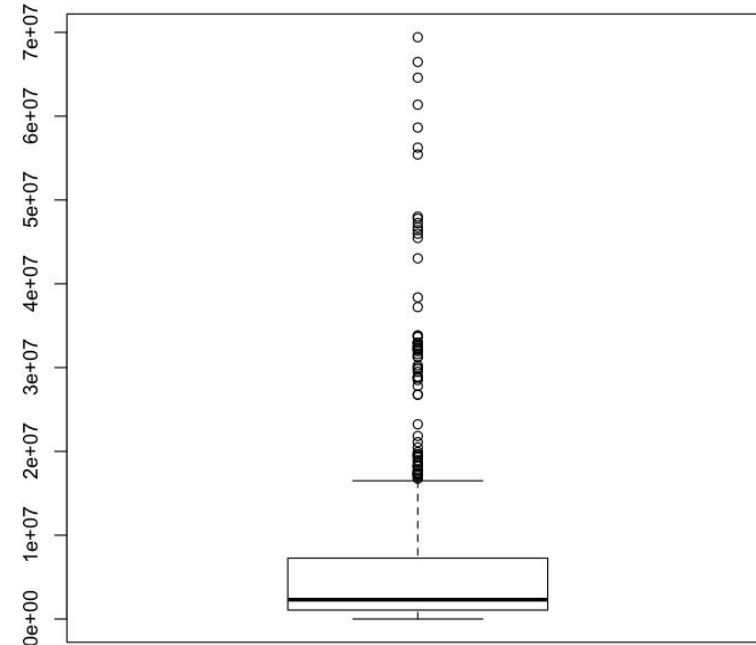
- now save your result locally as **csv** for further analysis
- -> astronaut2.csv

[Link to the SPARQL query](#)

astronaut	rank	time	year
http://dbpedia.org/resource/Yuri_Shargin	0.253642	855000.0	1960
http://dbpedia.org/resource/Evgeny_Tarelkin	0.278269	1.24137e+07	1974
http://dbpedia.org/resource/Gennadi_Sarafanov	0.289494	173520.0	1942
http://dbpedia.org/resource/Oleg_Artemyev	0.302007	1.462e+07	1970
http://dbpedia.org/resource/James_Dutton_(astronaut)	0.312072	1.30602e+06	1968
http://dbpedia.org/resource/Sergei_Revin	0.337499	1.07995e+07	1966
http://dbpedia.org/resource/Ulrich_Walter	0.342097	862800.0	1954
http://dbpedia.org/resource/Satoshi_Furukawa	0.343698	1.44288e+07	1964
http://dbpedia.org/resource/Brian_Duffy_(astronaut)	0.361668	3.51942e+06	1953
http://dbpedia.org/resource/Andrei_Borisenko	0.366213	1.41901e+07	1964
http://dbpedia.org/resource/Konstantin_Kozeyev	0.370514	849540.0	1967
http://dbpedia.org/resource/Anatoli_Ivanishin	0.376353	1.42831e+07	1969
http://dbpedia.org/resource/Donald_E._Williams	0.415804	1.03524e+06	1942
http://dbpedia.org/resource/Kenneth_D._Cameron	0.418345	2.02386e+06	1949
http://dbpedia.org/resource/William_Frederick_Fisher	0.424354	613020.0	1946
http://dbpedia.org/resource/Oleg_Novitskiy	0.426141	1.24137e+07	1971
http://dbpedia.org/resource/Shannon_Walker	0.426199	1.41091e+07	1965
http://dbpedia.org/resource/Alexander_Misurkin	0.431466	1.43649e+07	1977
http://dbpedia.org/resource/Anatoli_Levchenko	0.441409	683880.0	1941
http://dbpedia.org/resource/Andrew_M._Allen	0.442399	3.19752e+06	1955
http://dbpedia.org/resource/Ernst_Messerschmid	0.443108	607440.0	1945
http://dbpedia.org/resource/Charles_Camarda	0.443187	1.20072e+06	1952
http://dbpedia.org/resource/John_Oliver_Creighton	0.44587	1.45584e+06	1943
http://dbpedia.org/resource/Vyacheslav_Zudov	0.446007	173160.0	1942
http://dbpedia.org/resource/Franco_Malerba	0.456635	688500.0	1946
http://dbpedia.org/resource/James_C._Adamson	0.466948	1.20366e+06	1946
http://dbpedia.org/resource/Robert_J._Cenker	0.467142	525780.0	1948
http://dbpedia.org/resource/Maurizio_Cheli	0.469003	1.35966e+06	1959
http://dbpedia.org/resource/Igor_Volk	0.473657	1.01964e+06	1937
http://dbpedia.org/resource/Georgy_Shonin	0.485006	427320.0	1935

- 2nd example: include **birthdate** and **Time in Space**

```
[> astronauts2 <- read.csv("astronaut2.csv", header=FALSE)
[> summary(astronauts2)
      V1      V2      V3
Min.   :0.2536 Min.   : 30 Min.   :1921
1st Qu.:0.7259 1st Qu.:1062720 1st Qu.:1942
Median :0.9760 Median :2305630 Median :1952
Mean   :1.4951 Mean   :7383416 Mean   :1950
3rd Qu.:1.4615 3rd Qu.:7262100 3rd Qu.:1959
Max.   :26.3922 Max.   :69413900 Max.   :1978
[> boxplot(astronauts2)
[> boxplot(astronauts2$V2)
```

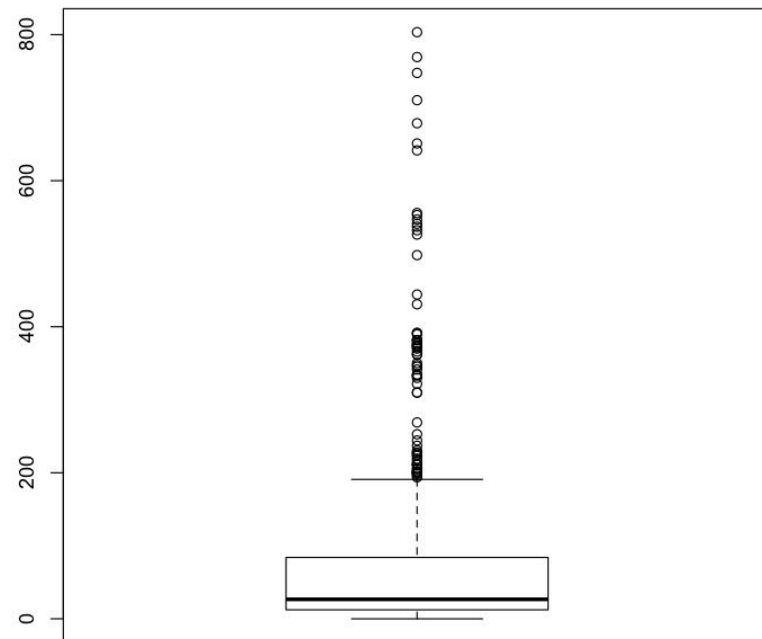


- 2nd example: include **birthdate** and **Time in Space**

```
[> summary(astronauts2$V2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   30  1063000  2306000  7383000  7262000 69410000

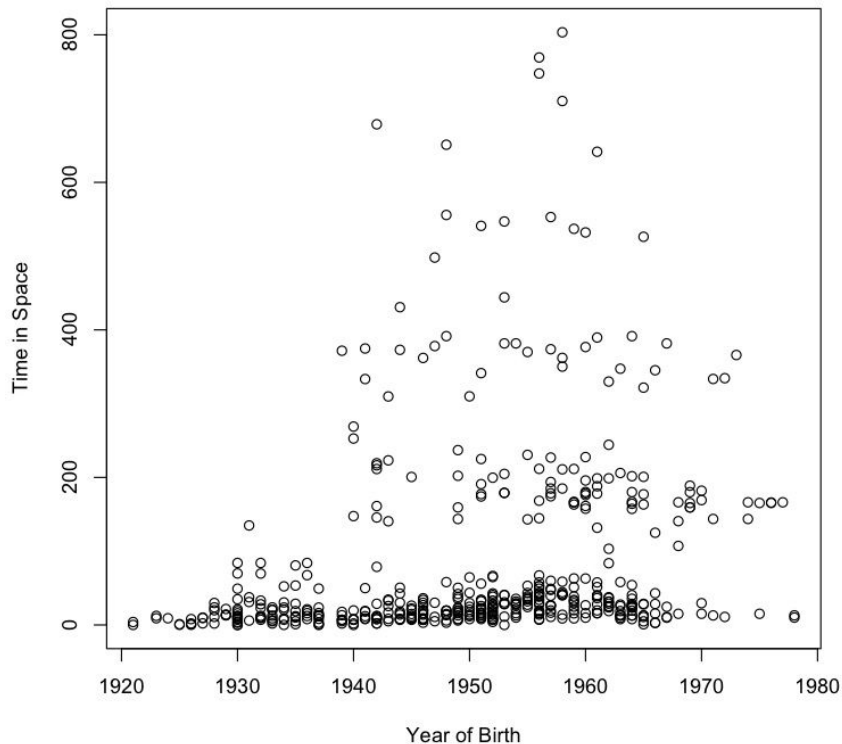
[> summary(astronauts2$V2/3600/24)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0003 12.3000  26.6900  85.4600  84.0500  803.4000

[> boxplot(astronauts2$V2/3600/24)
```



Linked Data Analytics

- 2nd example: include **birthdate** and **Time in Space**
 - **Hypothesis:** *“older astronauts have been in space for a longer time”*
 - ```
plot(astronauts2$V3, astronauts2$V2,
 ylab='Time in Space',
 xlab='Year of Birth')
```

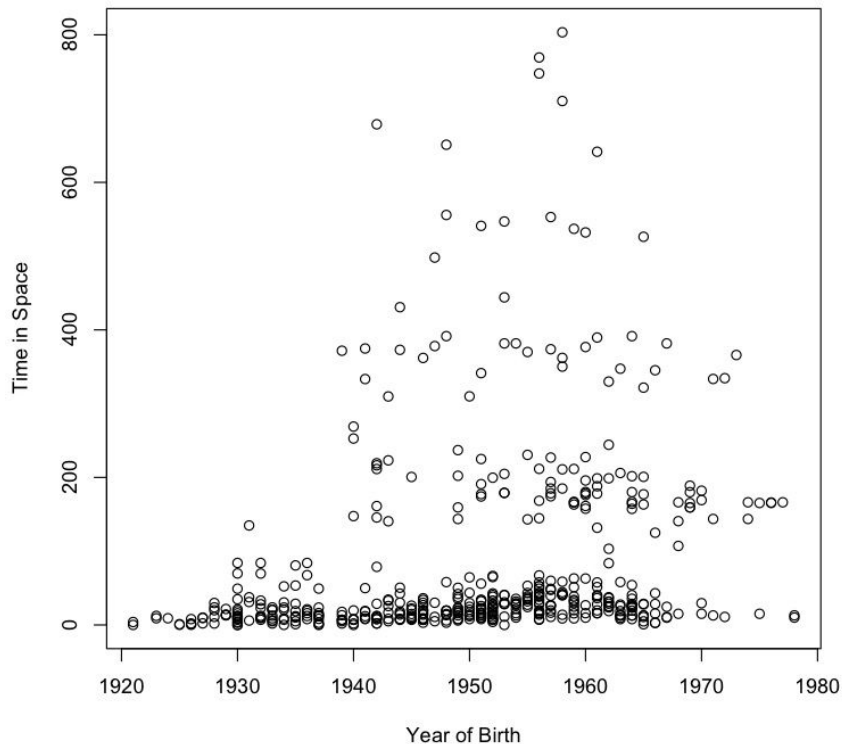


# Linked Data Analytics

- 2nd example: include **birthdate** and **Time in Space**
  - **Hypothesis:** *“older astronauts have been in space for a longer time”*

```
■ plot(astronauts2$V3, astronauts2$V2,
 ylab='Time in Space',
 xlab='Year of Birth')
```

```
■ cor(astronauts2$V2, astronauts2$V3)
[1] 0.2441805
```

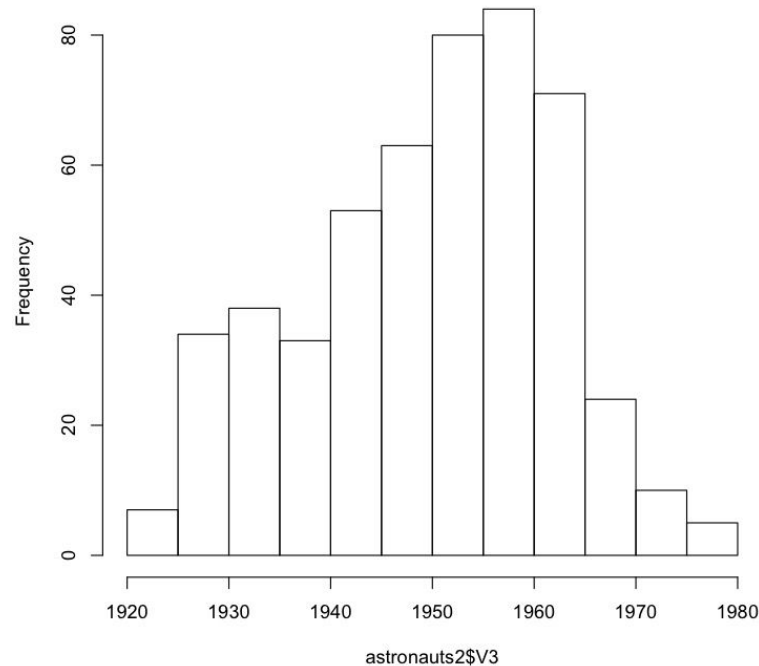


# Linked Data Analytics

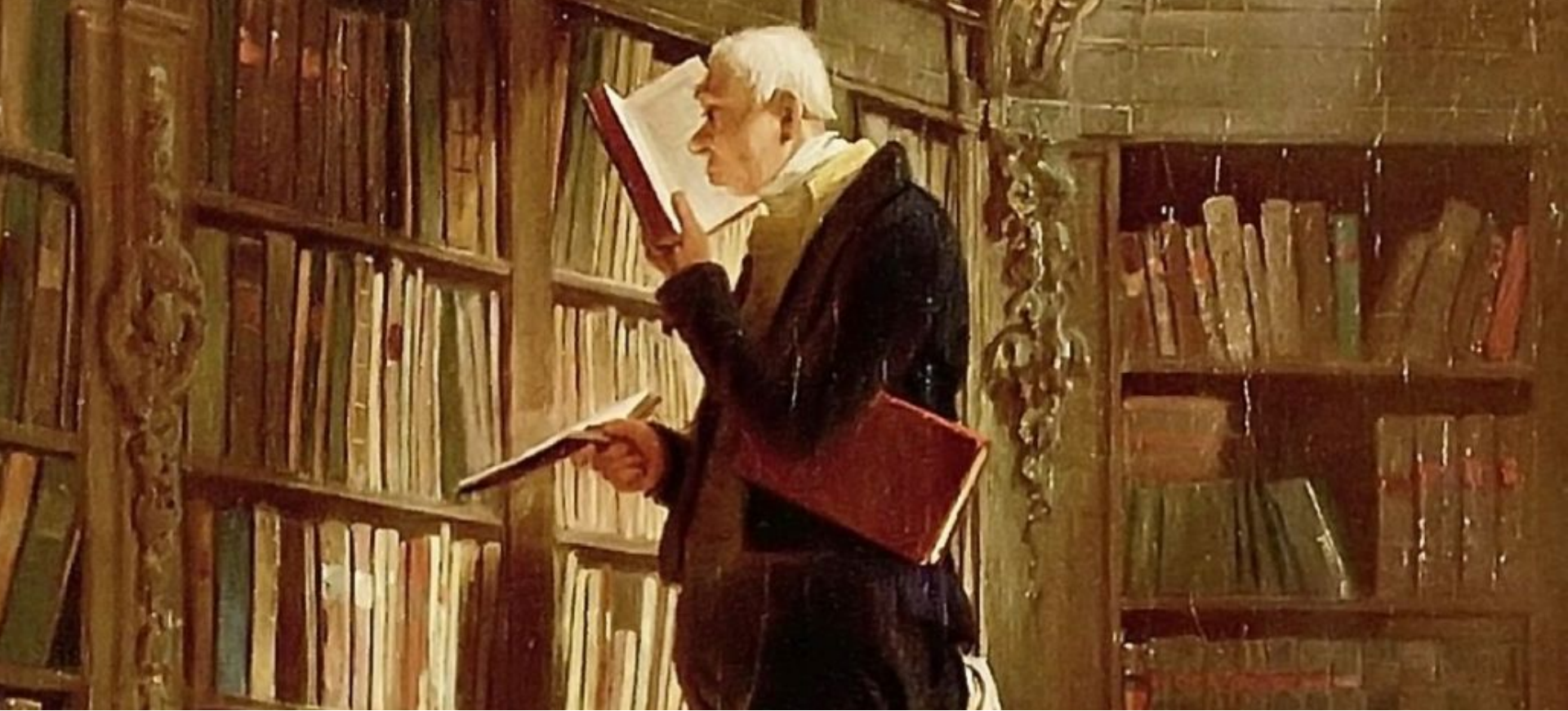
- 2nd example: include **birthdate** and **Time in Space**
  - **Question:** *“what’s the distribution of birthdates?”*

■ `hist(astronauts2$V3)`

Histogram of astronauts2\$V3



- Linked Data Analytics is an **iterative** and **exploratory** process
- **Visualization** is an important tool for data analysis
- Data analysis enables the **discovery of previous unknown relations**



## **10: EXTRA - Semantic Recommender Systems**

OpenHPI - Course Knowledge Engineering with Semantic Web Technologies

### **Lecture 6: Knowledge Engineering**